Bidirectional Attention Patterns with MASK Tokens

MLM Training Example:

Original: "The cat sits on the mat"

0.25   0.35   0.15   0.20   0.05

| The | [MASK] | sits | on | the | [MASK] |

Predicts: "cat"   Predicts: "mat"

pos 0   pos 1   pos 2   pos 3   pos 4   pos 5

Bidirectional Attention Matrix:
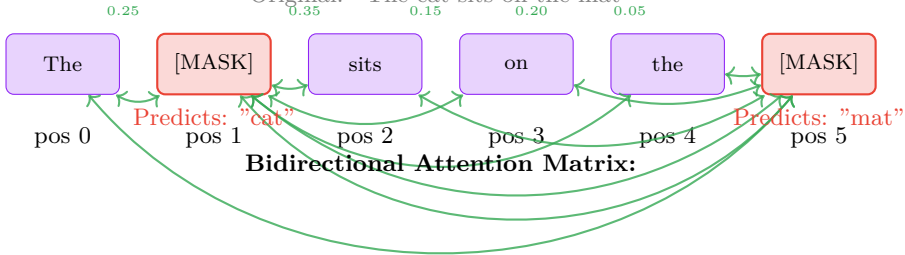
	The	[MASK]	sits	on	the	[MASK]
The						
[MASK]						

Key Differences from Autoregressive Models

Bidirectional Context: Uses both left & right context

No Causal Masking: Future tokens inform prediction

Parallel Prediction: Multiple masks predicted jointly

MLM Objective: Predict original tokens using full bidirectional context